DOCUMENT RESUME

ED 382 662 TM 023 097

AUTHOR Tang, K. Linda; And Others
TITLE The Effect of Small Calibration Sample Sizes on TOEFL IRT-Based Equating.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-59; TOEFL-TR-7
PUB DATE Dec 93
NOTE 52p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; *Equated Scores; *Estimation (Mathematics); *Item Response Theory; Pretests Posttests; *Sample Size; *Scaling; Simulation; Test Construction
IDENTIFIERS *BILOG Computer Program; Calibration; Item Parameters; LOGIST Computer Program; Test of English as a Foreign Language; Three Parameter Model

ABSTRACT

This study compared the performance of the LOGIST and BILOG computer programs on item response theory (IRT) based scaling and equating for the Test of English as a Foreign Language (TOEFL) using real and simulated data and two calibration structures. Applications of IRT for the TOEFL program are based on the three-parameter logistic (3PL) model. The results of the study show that item parameter estimates obtained from the smaller real data sample sizes were more consistent with the larger sample estimates when based on BILOG than when based on LOGIST. In addition, the root mean squared error statistics suggest that the BILOG estimates for the item parameters and item characteristic curves were closer in magnitude to the "true" parameter values than were the LOGIST estimates. The equating results based on the parameter estimates suggest that the rule of thumb recommendation that pretest sample sizes be at least 1,000 for LOGIST should be retained if at all possible. Eight tables and 13 figures present results of the analyses. Two appendixes contain specifications and summary statistics. (Contains 15 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 382 662

TOEFL®

December 1993

Technical Report

TR- 7

The Effect of Small Calibration Sample Sizes on TOEFL IRT-Based Equating

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

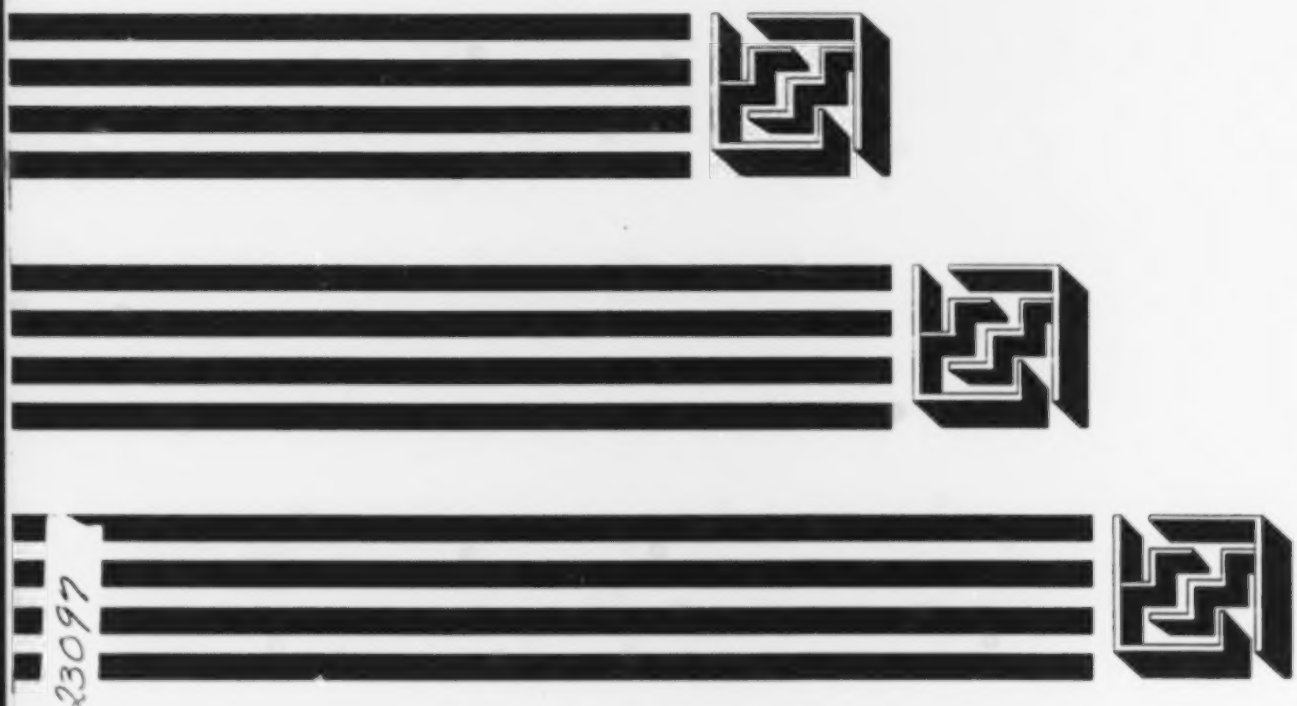
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

K. Linda Tang
Walter D. Way
Patricia A. Carey



TM023097

BEST COPY AVAILABLE

The Effect of Small Calibration Sample Sizes on TOEFL IRT-Based Equating

K. Linda Tang
Walter D. Way
Patricia A. Carey

Educational Testing Service
Princeton, New Jersey

RR-93-59

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Copyright © 1993 by Educational Testing Service. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both US and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, TOEFL, the TOEFL logo, and LOGIST are registered trademarks of Educational Testing Service.

Abstract

The present study compared the performance of LOGIST and BILOG on TOEFL IRT-based scaling and equating using both real and simulated data and two calibration structures. Applications of IRT for the TOEFL program are based on the three-parameter logistic (3PL) model.

The results of the study show that item parameter estimates obtained from the smaller real data sample sizes were more consistent with the larger sample estimates when based on BILOG than when based on LOGIST. In addition, the root mean squared error statistics suggest that the BILOG estimates for the item parameters and item characteristic curves were closer in magnitude to the "true" parameter values than were the LOGIST estimates.

The equating results based on the parameter estimates suggest that the rule of thumb recommendation that pretest sample sizes be at least 1000 for LOGIST should be retained if at all possible.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and, in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1992-93) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins (Chair)	Southern Illinois University at Carbondale
Linda Schinke-Llano	Millikin University
John Upshur	Concordia University

Acknowledgments

The authors thank Marna Golub-Smith for reviewing earlier drafts of this manuscript and for providing valuable comments. The authors also thank Kentaro Yamamoto for advising on BILOG parameter selections.

Table of Contents

	Page
Introduction	1
Data	3
Structure of the Data Sets	3
Sample Sizes	5
Methodology	6
Simulation I	6
Simulation II	6
Real Data	7
Data Analysis Methods	7
Results	9
Item Parameter Recovery	9
True Score Equating	15
Discussion and Conclusions	17
Figures	19
Appendix A	31
Appendix B	33
References	37

List of Appendices

Page

Appendix A LOGIST and BILOG Specifications 31

Appendix B Summary Statistics for the Data Sets Used in the Study 33

List of Figures

Figure 1 Calibration Design for Real Data Analyses and Simulation I 4

Figure 2 Calibration Design for Simulation II 4

Figure 3 Plots of Real Data A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=1000 Runs Vs. the N=4000 Runs 19

Figure 4 Plots of Real Data A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=2000 Runs Vs. the N=4000 Runs 20

Figure 5 Plots of Real Data A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=3000 Runs Vs. the N=4000 Runs 21

Figure 6 Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=1000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 22

Figure 7 Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=2000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 23

Figure 8 Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=3000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 24

Figure 9 Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=4000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 25

List of Figures (continued)

Figure 10 Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=2000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 26

Figure 11 Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=4000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 27

Figure 12 Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=6000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 28

Figure 13 Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=8000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters) 29

List of Tables

		Page
Table 1	Sample Sizes Used in the Real Data, Simulation I, and Simulation II	5
Table 2	Correlations between the Parameters Estimated by the Largest and Each of the Smaller Samples - Real Data Case	9
Table 3	Root Mean Squared Errors of the Estimated Parameters and ICCs - Real Data	11
Table 4	Correlations between the Generating Parameters and the Estimated Parameters - Simulation I	12
Table 5	Root Mean Squared Errors of the Estimated Parameters and ICCs - Simulation I	13
Table 6	Correlations between the Generating Parameters and the Estimated Parameters - Simulation II	14
Table 7	Root Mean Squared Error of the Estimated Parameters and ICCs - Simulation II	15
Table 8	Weighted Bias, Standard Deviation of the Difference, and Root Mean Squared Error in True Score Equating	16

Introduction

The IRT model of choice for the Test of English as a Foreign Language (TOEFL®) is the three-parameter logistic (or 3PL) model. In this model, the probability of a correct response for an item is a function of the examinee's ability and three item parameters. The 3PL model is expressed as:

$$P(\theta_j;\ a_i,\ b_i,\ c_i) = c_i + (1 - c_i) / \{1 + \exp[-Da_i(\theta_j - b_i)]\}, \quad i = 1,\ldots,\textit{number of items } (n), \quad j = 1,\ldots,\textit{number of examinees } (N). \tag{1}$$

In equation (1), c_i is the pseudo-guessing parameter for item i, a_i is the item discrimination parameter for item i, b_i is the item difficulty parameter for item i, θ_j is the ability parameter for examinee j, and D is a constant assuming the value of 1.7 (which is employed to make the logistic curve closely approximate the normal ogive model).

For the TOEFL test, the parameters of equation (1) are estimated by LOGIST®, which is based on the joint maximum likelihood (JML) approach suggested by Birnbaum (1968) and described in detail in Lord (1980). LOGIST finds the values of item and examinee parameters that simultaneously maximize a modified version of the joint likelihood function (Lord, 1980):

$$L = \prod_{j=1}^{N} \prod_{i=1}^{n} P(a_i,b_i,c_i;\theta_j)^{v_{ij}} [1 - P(a_i,b_i,c_i;\theta_j)]^{1-v_{ij}} \tag{2}$$

where $v_{ij} = 1$ *if examinee j responds correctly for item i,*
$v_{ij} = 0$ *if examinee j responds incorrectly for item i,*
$v_{ij} = 1/$*(number of choices) if examinee j omits item i.*

The authors of LOGIST recommend minimum calibration samples consisting of at least 1000 examinees and at least 40 items (Wingersky, Patrick, & Lord, 1988). Meeting these recommendations can sometimes be difficult when item parameters are estimated using pretest data because all pretest items are typically not administered to all examinees. If new item types were introduced into the TOEFL test, there would be no existing pool of pretested items available. In this case, there may be a need to pretest an extremely large number of the new item types in order to built up sufficiently large item pools in a reasonably short period of time. As a result, the need to meet recommended calibration sample sizes might have to be weighed against conflicting needs.

BILOG has been suggested as a viable alternative to LOGIST when calibration sample sizes are limited (Mislevy & Stocking, 1989; Way, Twing, & Ansley, 1988). BILOG uses the marginal maximum likelihood (MML) approach and the Bayes marginal modal solution (maximum a posteriori (MAP)) (Bock & Aitkin, 1981; Mislevy, 1986; Mislevy & Bock, 1984; Mislevy & Bock, 1989):

$$L = \prod_{j=1}^{N} \int \prod_{i=1}^{n} P(a_i, b_i, c_i; \theta)^{v_{ij}} [1 - P(a_i, b_i, c_i; \theta)]^{1-v_{ij}} dG(\theta) \qquad (3)$$

Where v_{ij} is defined the same as (2).

In equation (3), examinee parameters are removed from the estimation problem entirely by assuming a particular structure for the distribution of ability in the examinee population $G(\theta)$, and this structure is used in the estimation of item parameters. In addition, BILOG assumes that the a-parameter has a lognormal distribution and the c-parameter has a beta distribution. The BILOG parameter estimates maximize the logarithm of the product of the likelihood and the assumed "prior" distributions.

Using simulated data, Mislevy and Stocking (1989) found that for 1500 examinees BILOG appeared to recover the generating "true" parameters better than LOGIST for a 15 item test. However, for the 45 item test, the results from the two programs were very similar. Recent simulation studies with BILOG have suggested that adequate estimates of 3PL item parameters may be obtained using sample sizes as small as 250 when tests consist of 25 items or more (Harwell & Janosky, 1991).

Applications of LOGIST to simulated 3PL data have suggested that reasonable scaling and equating results may be obtained with smaller sample sizes, even though individual item parameters were not well estimated. Using a sample size of 1000, Yen (1987) found that in terms of recovering the individual item parameters, BILOG usually was substantially more accurate than LOGIST. However, in estimating the true score equating relationship, which is a common application at ETS, the two programs were about equally accurate when the number of items in a test was greater than or equal to 20. In a simulation study based on the TOEFL operational equating design, Way and Reese (1991) found that although correlations between LOGIST 3PL item parameter estimates and generating "true" parameters tended to be affected by sample size, neither the quality of model-data fit nor the quality of simulated equatings appeared to be sensitive to sample size. In their study, sample sizes for equating set items ranged from 600 to 1,500, while operational items were based on sample sizes ranging from 2,400 to 6,000. Broch and McKinley (1991) also found that LOGIST produced satisfactory results when applied to an incomplete data matrix

consisting of 120 total items, where 80 of the items were administered in sets of 20 to unique samples of 500 examinees, and 40 of the items were administered in sets of 10 to overlapping samples of 1000. However, real data were not examined in the above three studies.

The purpose of the present study was to explore the effects of small sample sizes on TOEFL IRT-based scaling and equating. The study examined a calibration design which is consistent with TOEFL pretesting, using both simulated and real data. In addition, the study examined a design using simulated data in which pretest items are calibrated independently of operational items. Estimation of item parameters using the PC version of LOGIST and PC-BILOG3 (Mislevy & Bock, 1989)[1] were compared using both real and simulated data of varying sample sizes.

Data

Structure of the Data Sets

Two simulations were carried out in the study. The first simulation (Simulation I) applied the procedures used by Way and Reese (1991). That is, 3PL model item parameters estimated by LOGIST for the TOEFL April 1991 administration were used as the generating "true" item parameters for the simulated data. In addition, systematic samples were selected from the population of TOEFL April 1991 ability estimates and served as the generating "true" ability parameters for the simulated data.[2] The generating parameters came from Section III[3] of the TOEFL test. It was assumed that enough similarities exist between the three sections of the TOEFL that results based on one of the sections would be relevant for the other two sections. Support for this assumption can be seen in the simulation results reported by Way and Reese (1991). The structure of the data for Simulation I was identical to the structure of the operational TOEFL data that were analyzed in the real data portion of the study. Figure 1 depicts the structure of these data. In Figure 1, the items 1 to 58 are operational items and items 59 to 178 are pretest items. Note that all examinees were administered the operational items and one of the four pretests.

---

[1]Documentation for running PC-LOGIST can be found in Dygert (1989). Both the version of PC-LOGIST and PC-BILOG used in this study were research rather than commercially available versions of these programs. Permission to use the programs was granted by the program custodians and/or authors.

[2]It should be noted that using LOGIST parameter estimates to generate the simulated data may favor LOGIST over BILOG. However, for the simulations to be relevant to the TOEFL, it was desirable to generate the data consistent with the existing TOEFL scale, which happens to be LOGIST based. [See Mislevy and Stocking (1989) for a discussion of this problem].

[3]Vocabulary and Reading Comprehension

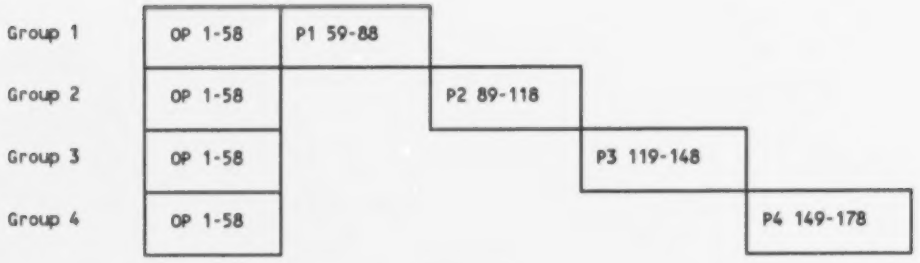


Figure 1
Calibration Design for Real Data Analyses and Simulation I
OP - Operational Items P - Pretest Items

For Simulation II, the data were structured as if eight thirty-item pretests were administered to eight different pretest samples, with 10 of the 30 items common to more than one pretest sample. The total number of items in this design is 200, with 40 "common" items and 160 unique items. This pretest design is depicted in Figure 2.

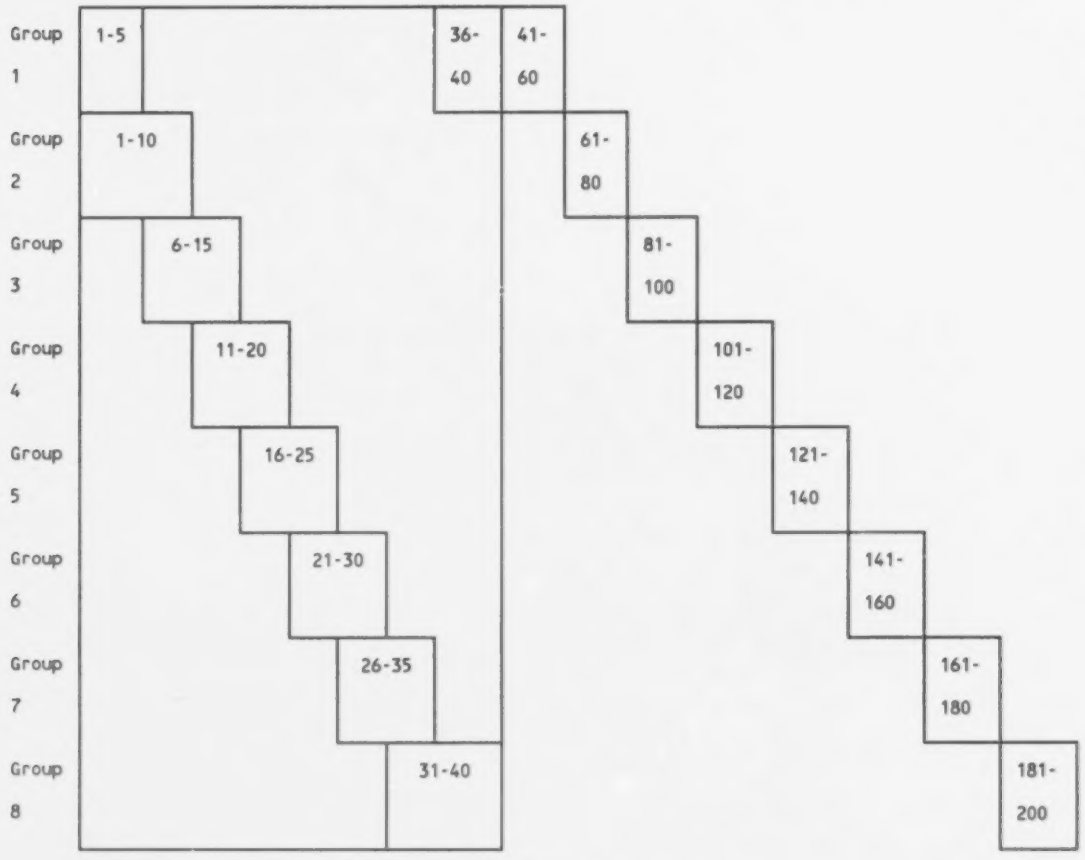


Figure 2
Calibration Design for Simulation II

The generating parameters for Simulation II consisted of the generating parameters for the four pretests used in Simulation I, as well as item parameter estimates from another operational TOEFL pretest administration. The generating ability estimates were the same as those used in Simulation I.

Sample Sizes

For Simulation I, Simulation II, and the Real Data Case, data sets were created using the following pretest sample sizes: 1000, 750, 500, and 250. For Simulation I and the Real Data Case, the sample sizes for the operational items were four times larger. For Simulation II, the sample sizes for the common items were two times larger. From Figures 1 and 2, it can be seen that the data were much sparser for Simulation II than for Simulation I or for the Real Data Case.

The sample sizes used in the present study are summarized in Table 1.

Table 1
Sample Sizes Used in the Real Data, Simulation I, and Simulation II

	Total N	N for Common Items	N for Unique Items
Simulation I/ Real Data Case	1,000	1,000	250
	2,000	2,000	500
	3,000	3,000	750
	4,000	4,000	1,000
Simulation II	2,000	500	250
	4,000	1,000	500
	6,000	1,500	750
	8,000	2,000	1,000

Methodology

Simulation I

The four data sets described in Table 1 and Figure 1 were generated on a microcomputer as follows: for each simulated item i and simulee j, a 0 or 1 response was assigned by comparing the probability of correct response as indicated by the 3PL model with the item i and person j parameter values to a random number drawn from a uniform(0,1) distribution. If the probability of correct response exceeded the value of the uniform random number, the item was scored as correct; otherwise, the item was scored as incorrect (Way & Reese, 1991)[4]. These data were calibrated using the PC version of LOGIST and BILOG3 (Mislevy & Bock, 1989) for a total of eight calibrations (four data sets X two estimation programs). The LOGIST and BILOG parameter specifications used in the calibrations are summarized in Appendix A[5]. The resulting item parameter estimates were then transformed to the TOEFL scale using a PC version of the Stocking and Lord (1983) procedure (TBLT). The common items used for the transformations were the 58 operational items. The generating parameter set was the "old form" and the estimated parameter sets were the "new forms" in the transformations.

Once the item parameter estimates for each of the eight calibrations were on a common scale, 58 items from the original 120 pretest items were selected to construct a typical Section III operational form. The "new" form was constructed such that the information function matched the information function for a typical TOEFL Section III. The new estimates of the 58 items from each of the eight data sets were equated to the "base form", the 58 items from the generating parameter set, using IRT true score equating procedure.

Simulation II

Data sets were simulated and calibrated, resulting in eight sets of item and ability parameter estimates. The item parameters were then transformed to the generating item parameter scale using TBLT, with the first 40 items as "common" items. The transformations were applied in this manner because there was no direct connection between the structure of the data in Simulation II and the current TOEFL equating design. Next, 58 items were selected from the 200 total items to serve as a "new" form. The new estimates from each of the eight data sets were equated to the "base form", the 58 items from the generating parameter set.

---

[4]Summary statistics for the Real Data Case, Simulation I, and Simulation II data sets are presented in Appendix B.

[5]The same parameter specifications were used in all the Simulation I, Simulation II, and Real Data Case calibration runs.

Real Data

For the Real Data Case, response data were obtained from an operational TOEFL pretest administration. For this portion of the study, responses to 58 Section III operational items and 120 pretest items were selected at random for samples of 4,000, 3,000, 2000, and 1,000 examinees. For each of these samples, the number of responses to the pretest item sets were 1,000, 750, 500, and 250, respectively.

Each of the four data sets was calibrated using PC-LOGIST and PC-BILOG3. For this portion of the study, within each estimation program, the calibrations based on 4,000 examinees taking operational items and 1,000 examinees taking pretest items, were used as the standard to compare the results of the calibrations carried out using the smaller data sets. Thus, the results of the LOGIST and BILOG runs were only indirectly comparable. The item parameter estimates for the three sets of smaller sample sizes were transformed to the scale of the calibrations based on the largest sample size using TBLT. As in the case of Simulation I, 58 of the 120 pretest items were selected to form a "typical" section III form. Six true score equatings were then carried out. That is, the three sets of PC-BILOG estimates for the smaller sample sizes were equated to the PC-BILOG estimates obtained from the largest data set. Similarly, three sets of PC-LOGIST estimates for the smaller sample sizes were equated to the PC-LOGIST results for the largest data set in the same manner.

Data Analysis Methods

The LOGIST and BILOG results were compared in two ways: item parameter recovery and true score equating.

Comparison of LOGIST and BILOG in Terms of Parameter Recovery. The correlations between the generating true parameters and the parameter estimates produced by the two programs were computed for the common items and the unique items. The root mean squared error, which was defined as

$$\left[\frac{\sum_i (\textit{Estimated Parameter}_i - \textit{"True" Parameter}_i)^2}{\textit{Number of items}}\right]^{1/2}, \qquad (4)$$

and the root mean squared error between Item Characteristic Curves (ICCs), which was defined as

$$\left[\frac{1}{\textit{Number of items}}\sum_i \frac{1}{31}\sum_{j=1}^{31}\left[P(\theta_j;\hat{a}_i,\hat{b}_i,\hat{c}_i) - P_{TRUE}(\theta_j;a_i,b_i,c_i)\right]^2\right]^{1/2} \qquad (5)$$
$$\textit{where } \theta_j = -3.0 + 0.2(j-1),\ j = 1,\ldots,31$$

were also computed for the common items and the unique items. In these equations, "true parameter" and P_{TRUE} refer to the generating parameters or probabilities obtained by the generating parameters in Simulations I and II, and to the parameters estimated by the largest sample (N=4000) or the probabilities obtained by the same sample in the Real Data Case. In addition, bivariate plots of the parameter estimates versus the generating parameters were produced for Simulations I and II. In the Real Data Case, similar plots were produced using the parameters estimated by the largest sample in place of the generating parameters.

Comparison of LOGIST and BILOG in Terms of True Score Equating. The weighted differences between each of the 59 estimated true scores and the corresponding "base form" true scores were computed, where the weights were obtained according to the following procedure:

(1) Use the ability parameters from the N=1000 simulation condition and the item parameters for the 58 selected items to generate a 1000 X 58 matrix of 0 or 1 item responses.

(2) Sum the item responses for each simulee to obtain a raw score.

(3) Calculate weights by dividing the frequency of each raw score by 1000.

In addition, the standard deviation of the weighted difference; the weighted bias, which was defined as the weighted mean difference; and the weighted root mean squared error, which was defined as

$$\left[\frac{1}{59}\sum_{i=1}^{59}(\textit{Estimated True Score}_i - \textit{"Base Form" True Score}_i)^2 \times \textit{Weight}_i\right]^{1/2}, \qquad (6)$$

were computed. Note that the square of the weighted bias plus the square of the standard deviation of the weighted difference is equal to the square of the weighted Root Mean Squared Error. The true score differences for each data set were also plotted.

Results

Item Parameter Recovery

Real data case. The correlations between the parameters estimated using the largest and each of the smaller samples for the Real Data Case are presented in Table 2. The bivariate plots of the parameter estimates produced by LOGIST and BILOG are presented in Figures 3 through 5[6].

Table 2
Correlations between the Parameters Estimated by the Largest and Each of the Smaller Samples
Real Data Case

Item Parameters	A LOGIST/BILOG	B LOGIST/BILOG	C LOGIST/BILOG
Operational Items (58) Sample Sizes			
1000 vs. 4000	0.77/0.90	0.94/0.98	0.39/0.69
2000 vs. 4000	0.81/0.84	0.98/0.98	0.70/0.75
3000 vs. 4000	0.92/0.93	0.99/0.99	0.83/0.80
Pretest Items (120) Sample Sizes			
250 vs. 1000	0.36/0.67	0.89/0.97	0.22/0.55
500 vs. 1000	0.57/0.78	0.94/0.98	0.34/0.63
750 vs. 1000	0.63/0.75	0.96/0.98	0.51/0.60

Table 2 shows that for both LOGIST and BILOG, the correlations increased when the sample size used in the parameter estimations increased; except for the a-estimates based on BILOG, which did not consistently increase with sample size.

[6]In these plots, the operational items (common items in Simulation II) and the pretest items (unique items in Simulation II) were overlaid: '*' represents an operational item (a common item in Simulation II) and 'o' represents a pretest item (a unique item in Simulation II).

In general, the correlations between the BILOG estimates were identical or higher than the correlations between the LOGIST estimates for all three parameter estimates and across all experimental conditions with one exception: c-estimates obtained when the sample size was 3000.

For b-estimates, the correlations based on BILOG were only sightly higher than those based on LOGIST. The maximum difference between the correlations based on LOGIST and those based on BILOG was 0.08 for b-estimates (N=250).

For a-estimates, the correlations based on BILOG were much higher than those based on LOGIST when the sample size was 1000 or smaller. The lowest correlation between a-estimates was 0.67 for BILOG and 0.36 for LOGIST, which occurred for the pretest items when the sample size was the smallest (N=250). The low correlation between the a-estimates for LOGIST illustrated in Figure 3 is indicative of the differences in estimation approach taken by BILOG and LOGIST. While BILOG utilizes a lognormal prior distribution to constrain the a-parameters, LOGIST uses the data available, which decreases with sample size. LOGIST then utilizes AMAX to restrict the a-parameters for poor items. The items that were set to AMAX are represented by the horizontal line in Figure 3.

The correlations between the c-estimates were lower than those for the a- and b-estimates for both LOGIST and BILOG. This is not surprising because at the lower end of the ability range fewer examinees were available to estimate the lower asymptote of the ICC. However, the correlations between the c-estimates were much lower for LOGIST than for BILOG when sample sizes were 1000 or smaller. The magnitudes of the correlation coefficients reflect the different approaches taken by LOGIST and BILOG in c-parameter estimation. For LOGIST, the items that contain little information about their lower asymptotes ($b-2/a < -2.5$) were pooled to provide a common estimate. In Figures 3 to 5, it can be seen that these common c-parameter estimates formed a vertical or horizontal straight line, which indicated that these common c-parameter estimates did not correlate in the two estimations. For BILOG, on the other hand, the estimation was influenced by the likelihood function and the prior distribution. The contribution from the likelihood function increases with sample size. The contribution from the prior remains constant with respect to sample size. Because the c-estimates from the largest sample were based on the same prior distribution as from the smaller sample, the correlation between the two sets of estimates was higher compared with that based on LOGIST. When the overall sample size increased to 2000, the magnitudes of the correlation coefficients based on the two programs are similar.

Table 3 presents the root mean squared errors of the parameter estimates (RMSEE). For LOGIST, the RMSEEs decreased as the sample size increased. For BILOG, the same trend was observed for b- and c-estimates and for the ICCs. However, for a-estimates, the RMSEE was 0.015 larger for sample size 750 compared with that for sample size 500, and was 0.027 larger for sample size 2000 compared with that for sample size 1000. In addition, the decrease in RMSEEs of a-estimates for LOGIST was 2.43 times greater than that for

BILOG as sample size increased from 250 to 3000. Therefore, the increase in sample size did not have a strong impact on the RMSEEs of the a-estimates for BILOG as it did for LOGIST. The fluctuation of the RMSEE for BILOG a-estimates was consistent with the fluctuation of the correlation coefficients for BILOG presented in Table 2.

Table 3
Root Mean Squared Errors of the Estimated Parameters and ICCs
Real Data

Item Parameters	A LOGIST/BILOG	B LOGIST/BILOG	C LOGIST/BILOG	ICC LOGIST/BILOG
Operational Items (58)				
Sample Sizes				
1000	0.213/0.113	0.260/0.149	0.079/0.052	0.029/0.024
2000	0.168/0.140	0.147/0.143	0.051/0.047	0.022/0.021
3000	0.103/0.088	0.113/0.112	0.038/0.039	0.019/0.019
Pretest Items (120)				
Sample Sizes				
250	0.424/0.220	0.532/0.263	0.141/0.047	0.060/0.039
500	0.326/0.181	0.405/0.218	0.110/0.042	0.047/0.033
750	0.307/0.196	0.338/0.197	0.104/0.038	0.043/0.032

The RMSEEs of the parameter estimates for BILOG were smaller than those for LOGIST in all experimental conditions, except for c-estimates obtained from sample size 3000. When the sample size reached 2000, the differences in RMSEE between BILOG and LOGIST were smaller than 0.01 for b- and c-parameter estimates, and smaller than 0.03 for a-parameter estimates.

The root mean squared errors for the ICCs (RMSEICC) are also presented in Table 3. Again, this statistic decreased as sample size increased for both LOGIST and BILOG, and the values of the statistic were smaller or identical for BILOG than those for LOGIST for all experimental conditions. When sample size was 250, the RMSEICC for LOGIST was 1.5 times that for BILOG. However, when sample size increased to 2000, the difference between the RMSEICCs of LOGIST and BILOG was less than 0.001.

The correlation coefficients, bivariate plots, RMSEEs, and RMSEICCs obtained form the Real Data Case show that BILOG estimates obtained from smaller samples ($N < 2000$) were more consistent with those obtained from the larger sample ($N = 4000$) than the LOGIST estimates. When the sample size reached 2000, the differences in the consistency of the parameter estimates between the two programs were trivial.

Simulation I. The correlations between the parameters estimated using each of the four simulated samples and the generating "true" parameters are presented in Table 4. The bivariate plots of the parameter estimates produced by LOGIST and BILOG against the generating parameters are presented in Figures 6 through 9.

For both LOGIST and BILOG, the correlations between each of the estimated a- and b-parameters and the generating parameters increased or remained identical as sample size increased in most cases. The exception worthy of notice was that the correlation between the LOGIST a-estimates and the a-parameters was 0.04 lower for pretest sample size 750 than that for pretest sample size 500. This inconsistency may be explained by examining Figure 8: there is one estimate which was set to AMAX, however, the corresponding generating parameter value was about 0.3.

Table 4
Correlations between the Generating Parameters and the Estimated Parameters
Simulation I

Item Parameters	A LOGIST/BILOG	B LOGIST/BILOG	C LOGIST/BILOG
Operational Items (58)			
Sample Sizes			
1000	0.86/0.87	0.95/0.96	0.57/0.68
2000	0.90/0.91	0.97/0.98	0.69/0.79
3000	0.90/0.92	0.95/0.97	0.63/0.84
4000	0.95/0.96	0.96/0.99	0.61/0.86
Pretest Items (120)			
Sample Sizes			
250	0.60/0.59	0.93/0.94	0.51/0.49
500	0.73/0.67	0.95/0.96	0.70/0.70
750	0.69/0.67	0.96/0.96	0.73/0.64
1000	0.78/0.77	0.98/0.98	0.66/0.74

Comparing LOGIST with BILOG, the magnitudes of the correlation coefficients were almost identical for the a- and b-estimates across all experimental conditions. Thus, the tendency for LOGIST item parameter estimates to fluctuate when sample sizes were less than 1000, which was seen in the Real Data Case, did not seem to be associated with similar fluctuations in the correlations between the estimates and the generating item parameters in Simulation I. Table 5 presents the RMSEEs and the RMSEICCs, and it indicates that

with one exception, (the c values for the pretest sample size of 750), all of these statistics were lower for the BILOG estimates than for the LOGIST estimates. This result suggests that BILOG estimates are slightly closer in magnitude to the "true" parameters than are LOGIST estimates, and that BILOG estimations are more consistent than LOGIST across sample size. However, the differences were small.

Table 5
Root Mean Squared Errors of the Estimated Parameters and ICCs
Simulation I

Item Parameters	A LOGIST/BILOG	B LOGIST/BILOG	C LOGIST/BILOG	ICC LOGIST/BILOG
Operational Items (58)				
Sample Sizes				
1000	0.158/0.131	0.248/0.215	0.077/0.069	0.027/0.024
2000	0.119/0.104	0.193/0.149	0.068/0.058	0.023/0.021
3000	0.124/0.097	0.223/0.171	0.073/0.051	0.019/0.015
4000	0.086/0.069	0.216/0.117	0.073/0.047	0.017/0.014
Pretest Items (120)				
Sample Sizes				
250	0.337/0.264	0.428/0.399	0.120/0.093	0.057/0.049
500	0.246/0.233	0.349/0.324	0.088/0.083	0.038/0.035
750	0.270/0.238	0.317/0.302	0.077/0.082	0.033/0.031
1000	0.214/0.190	0.238/0.237	0.085/0.072	0.029/0.027

The results evaluated by the correlations, the RMSEEs, and the RMSEICCS showed that LOGIST and BILOG performed about equally well in parameter recovery for Simulation I. The improved performance for LOGIST in Simulation I over the Real Data Case indicates that LOGIST estimates are more consistent with the "true" parameters than with the parameters estimated by a larger sample.

Simulation II. The correlations between the parameters estimated using each of the four simulated samples and the generating "true" parameters are presented in Table 6. The bivariate plots of the parameter estimates produced by LOGIST and BILOG against the generating parameters are presented in Figures 10 through 13.

Table 6
Correlations between the Generating Parameters and the Estimated Parameters
Simulation II

Item Parameters	A LOGIST/BILOG	B LOGIST/BILOG	C LOGIST/BILOG
Common Items (40)			
Sample Sizes			
500	0.72/0.60	0.95/0.97	0.43/0.50
1000	0.67/0.67	0.98/0.98	0.61/0.61
1500	0.84/0.83	0.99/0.99	0.80/0.82
2000	0.84/0.90	0.95/0.98	0.65/0.76
Unique Items (160)			
Sample Sizes			
250	0.53/0.50	0.93/0.95	0.53/0.54
500	0.69/0.65	0.90/0.97	0.68/0.63
750	0.80/0.76	0.96/0.97	0.69/0.72
1000	0.81/0.76	0.97/0.98	0.75/0.75

The correlations between the generating b-parameters and the estimated b-parameters for BILOG were slightly higher or identical compared with those for LOGIST. However, the differences were trivial: the maximum difference in the correlation coefficients obtained from the two programs was 0.07 (N=500). The correlations between the generating c-parameters and the estimated c-parameters for BILOG were higher than those for LOGIST in five experimental conditions, identical in two experimental conditions and 0.05 lower in one experimental condition (N=500).

For the a-parameter, the correlations for LOGIST were slightly higher or identical compared with those for BILOG, except when the sample size was 2000. Examining the bivariate plots of a-parameters based on BILOG in Figures 10 through 13, it can be seen that several a-parameters that were greater than 1.5 in the generating parameter set had estimated values less than 1.25. One property of the BILOG estimates is that by using a prior distribution, the estimates tend to shrink toward the estimated mean when the sample size is small. Because the lognormal prior distribution was not imposed in the generating parameters (the original LOGIST estimates), the shrinkage toward the estimated parameter mean of the BILOG estimates decreased their correlation with the generating parameters, given that the correlation is the sum of the products of the difference between each estimate and the corresponding mean estimate. Because the sample sizes used in Simulation II were smaller than those in the Real Data Case and in Simulation I, the shrinkage toward the

estimated mean effect of BILOG was more apparent than in the Real Data Case and in Simulation I.

In terms of the RMSEEs and the RMSEICCs, as presented in Table 7, BILOG still performed slightly better than LOGIST on almost all experimental conditions in Simulation II. These statistics are less sensitive to the shrinkage toward the mean effect of BILOG. The slightly better performance of BILOG over LOGIST evaluated by the RMSEEs and the correlations of b- and c- parameters may be more convincing when taking into account that the generating parameters were LOGIST estimates.

Table 7
Root Mean Squared Error of the Estimated Parameters and ICCs
Simulation II

Item Parameters	A LOGIST/BILOG	B LOGIST/BILOG	C LOGIST/BILOG	ICC LOGIST/BILOG
Common Items (40)				
Sample Sizes				
500	0.229/0.275	0.421/0.277	0.102/0.092	0.051/0.041
1000	0.238/0.240	0.287/0.219	0.087/0.085	0.042/0.034
1500	0.209/0.185	0.203/0.177	0.063/0.072	0.030/0.025
2000	0.174/0.136	0.385/0.203	0.079/0.073	0.033/0.024
Unique Items (160)				
Sample Sizes				
250	0.370/0.315	0.457/0.316	0.116/0.092	0.061/0.048
500	0.278/0.253	0.631/0.256	0.080/0.082	0.046/0.037
750	0.273/0.220	0.326/0.223	0.079/0.075	0.040/0.032
1000	0.238/0.206	0.282/0.215	0.068/0.073	0.036/0.030

True Score Equating

The weighted bias, weighted standard deviation of the difference (SD Diff), and the weighted root mean squared errors in true score equating (RMSEEQ) of the Real Data Case, Simulation I, and Simulation II are presented in Table 8. In addition, the equating differences are also plotted in Figures 14 through 16.

Table 8
Weighted Bias, Standard Deviation of the Difference, and Root Mean Squared Error In True Score Equating

	Bias LOGIST/BILOG	SD Diff LOGIST/BILOG	RMSE LOGIST/BILOG
Real Data			
Sample Sizes			
250 (N=1000)	0.276/0.193	0.538/0.629	0.604/0.658
500 (N=2000)	0.254/0.222	0.684/0.713	0.730/0.746
750 (N=3000)	0.498/0.463	0.439/0.465	0.664/0.656
Simulation I			
Sample Sizes			
250 (N=1000)	-0.284/-0.352	0.476/0.552	0.554/0.655
500 (N=2000)	-0.001/0.015	0.425/0.460	0.425/0.460
750 (N=3000)	0.358/0.347	0.362/0.342	0.509/0.487
1000 (N=4000)	-0.175/-0.139	0.297/0.103	0.345/0.173
Simulation II[a]			
Sample Sizes			
250/500 (N=2000)	-0.405/-0.293	0.192/0.367	0.448/0.470
500/1000 (N=4000)	0.161/0.022	0.307/0.304	0.346/0.305
750/1500 (N=6000)	0.272/0.215	0.361/0.093	0.452/0.235
1000/2000 (N=8000)	0.093/-0.061	0.295/0.151	0.309/0.163

[a]Note: For Simulation II, the 58 equating items were selected from the 200 total items, which include 40 common items and 160 unique items. The sample size of "examinees" taking the common items is twice as large as that taking the unique items.

Table 8 shows that the biases for BILOG were smaller than those for LOGIST except in Simulation I when sample sizes were 250 and 500. However, the maximum difference in bias between LOGIST and BILOG was only 0.154 (Simulation II, N=8000). Because bias is the average of the differences between the true scores based on the estimated parameters and based on the generating parameters (parameters estimated by the largest samples in the Real Data Case), it is informative to look at the individual true score differences also, shown in Figures 14 through 16.

Figure 14 shows that the individual true score difference was less than two points for LOGIST, and slightly larger than two points for BILOG in the true score range from 10 to 16 when the sample size was 1000, and at true scores 15 and 16 when the sample size was 2000. For Simulation I, Figure 15 shows that the differences were less than two points for both LOGIST and BILOG, except when N=2000. In that case, true scores ranging from

11 to 14 had differences slightly greater than two points for LOGIST. In Simulation II, all the true score differences were less than two points. Furthermore, for BILOG, the true score differences were less than 0.5 in the true score range 20 to 46, i.e., in the ability (θ) range -1.78 to 0.49.

One phenomenon observed consistently for both LOGIST and BILOG across all the experimental conditions was that the largest true score difference occurred in the low score range, i.e., true scores less than 16, or in other words, ability (θ) less than -3.0. In practice, less than three percent of examinees score in that range.

For the SD Diff and RMSEEQ presented in Table 8, LOGIST had slightly smaller values in half of the experimental conditions, and BILOG had smaller values in the other half. One trend observed for BILOG was that the SD Diff and RMSEEQ tended to decrease as the sample size increased. No consistent pattern was observed for LOGIST, however, the RMSEEQ reached smallest value when the sample sizes were largest for both Simulations I and II.

Based on the results obtained from the Real Data, Simulation I, and Simulation II, it is concluded that the performance of LOGIST and BILOG parameter estimates provided similar equating results.

Discussion and Conclusions

Using real and simulated data, the present study compared the performance of LOGIST and BILOG on TOEFL IRT-based scaling and equating. The results of the study show that item parameter estimates based on the smaller real data sample sizes were more consistent with the larger sample estimates when based on BILOG than when based on LOGIST. In addition, RMSE statistics suggested that the BILOG estimates for the item parameters and ICCs were closer in magnitude to the "true" parameter values than were the LOGIST estimates. These findings are consistent with Mislevy and Stocking (1989) and Yen (1987). In terms of equating, the smaller sample sizes increased the equating RMSEs for both BILOG and LOGIST in both Simulations I and II, and the equating errors increased the most when the pretest sample size decreased from 1000 to 750. These results suggest that the rule of thumb recommendation that pretest sample sizes be at least 1000 should be retained if at all possible. Note that TOEFL volumes are sufficient to support sample sizes of 1000 given the current level of pretesting, even though the level of pretesting has increased drastically in recent years.

Because BILOG appears to provide some advantages over LOGIST in terms of item parameter estimation, this might suggest that BILOG should be considered if the TOEFL program pursues computer adaptive testing (CAT). In addition, if binary (scored right or wrong) items for a new TOEFL measure need to be calibrated, the calibration structure utilized in Simulation II should probably be used. Because implementation of a new

TOEFL measure would probably involve the establishment of a new IRT scale, BILOG could be utilized and would be preferred over LOGIST. However, the TOEFL program should continue to use LOGIST for the current test because the equating results of LOGIST and BILOG were similar when the sample sizes were larger than 1000, and also because switching to BILOG would cause a discontinuity in the current scale, since the relationship between LOGIST and BILOG parameter estimates is non-linear (Mislevy & Stocking, 1989).

Some limitations of the study are worth mentioning. Given that this study examined only one TOEFL section, the generalizability of the results to the other sections need to be investigated. In addition, the simulations used in the study involved only a single replication per cell, which may have masked trends that would be apparent with multiple replications. Another limitation of the study is that the choices for program parameter options, particularly for BILOG, may not have been optimal.

Based on the results of the current study, it might be interesting to carry out further research, such as the exploration of calibration procedures that allow for both binary and polytomous item responses; and scaling and scoring procedures that might be applied to a computerized version of the TOEFL test.

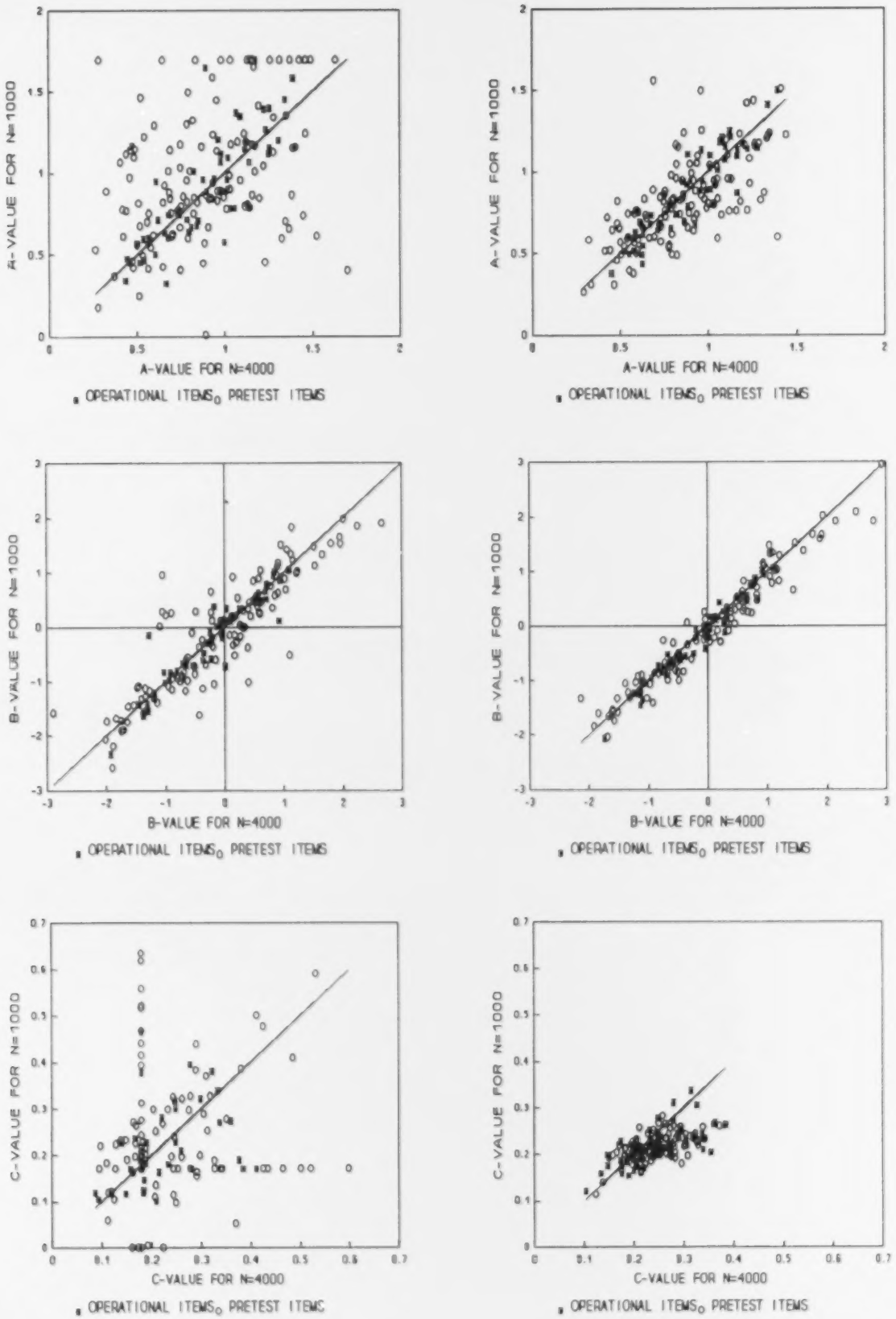


Figure 3: Plots of Real Data A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=1000 Runs Vs. the N=4000 Runs

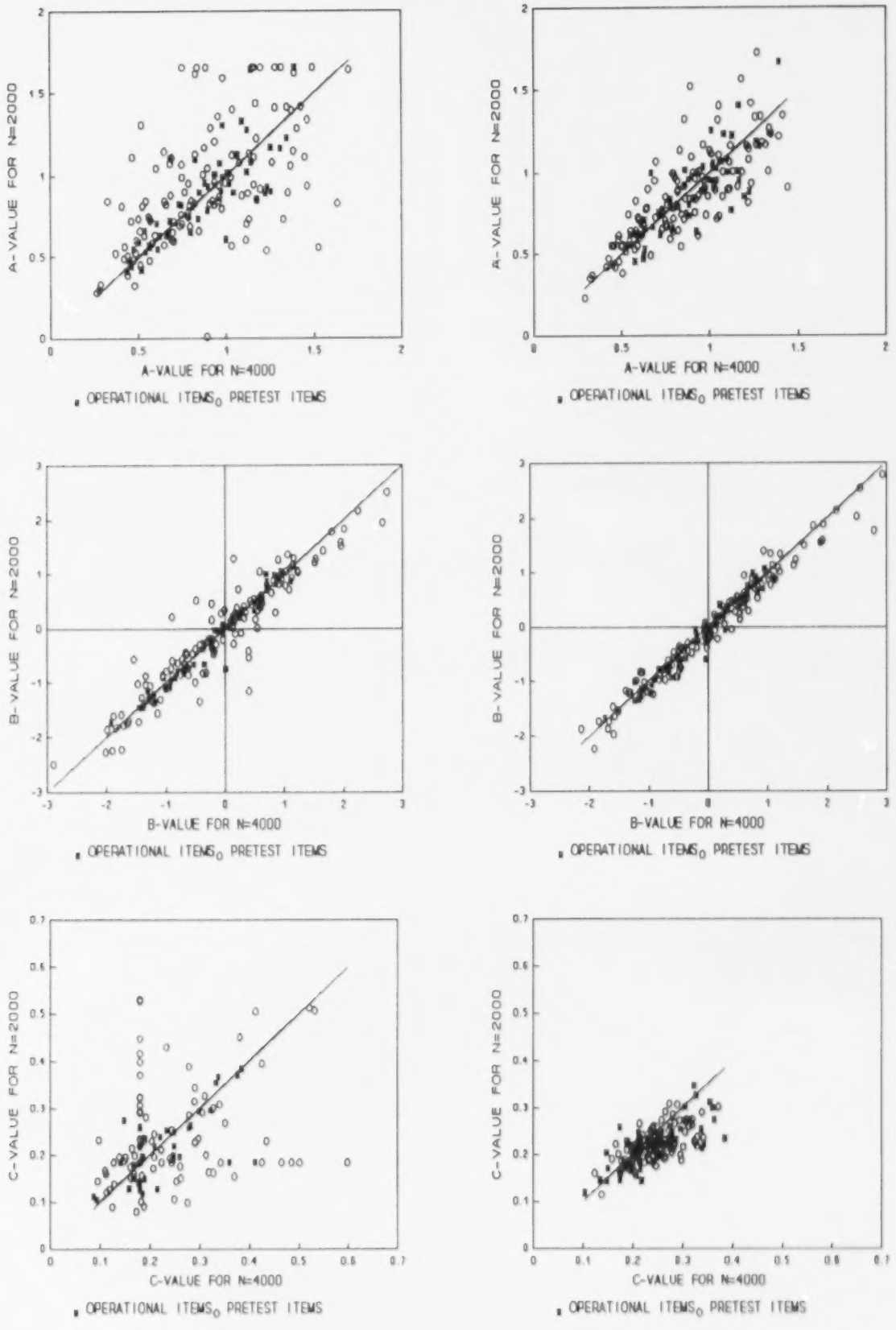


Figure 4: Plots of Real Data A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=2000 Runs Vs. the N=4000 Runs

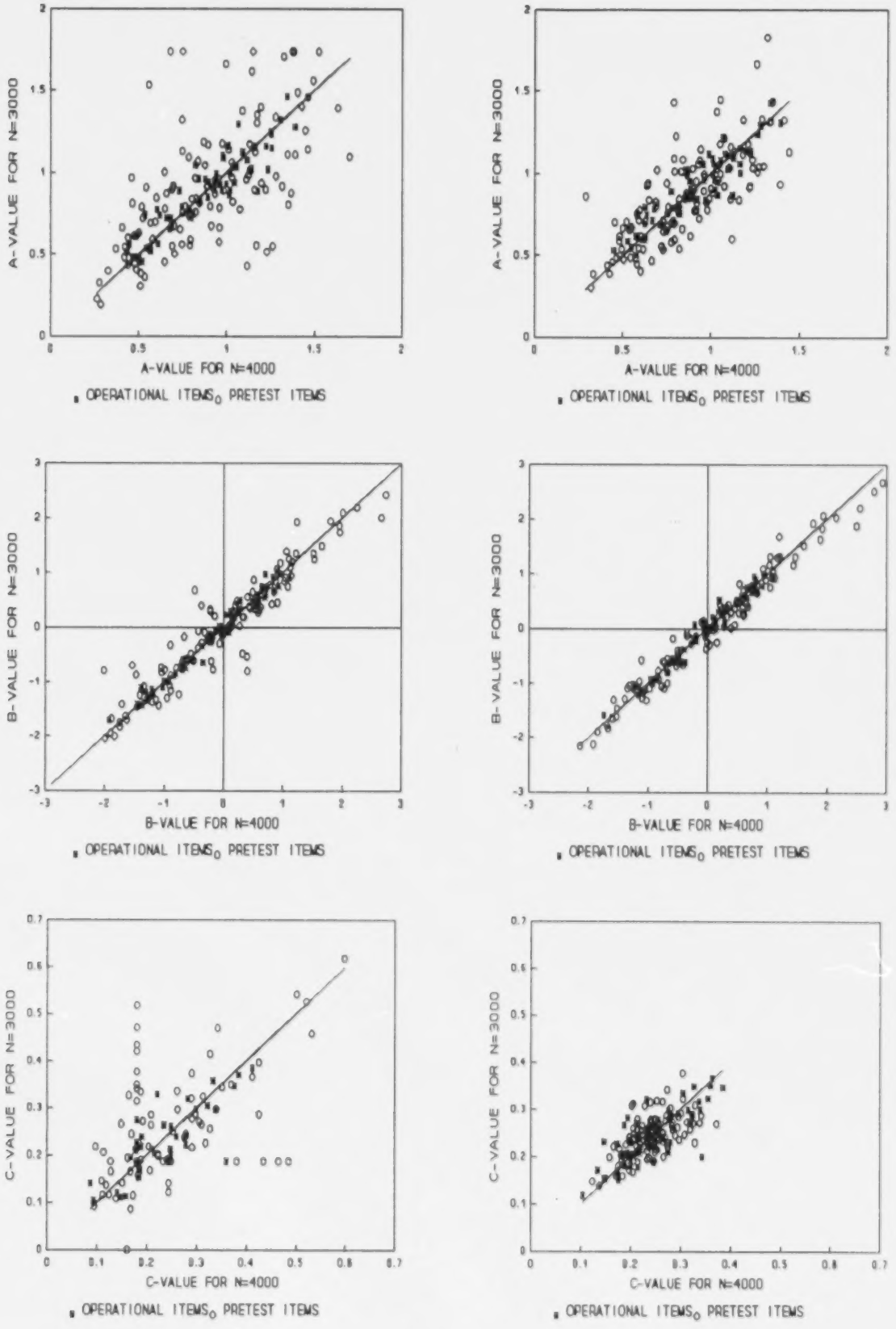


Figure 5: Plots of Real Data A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=3000 Runs Vs. the N=4000 Runs

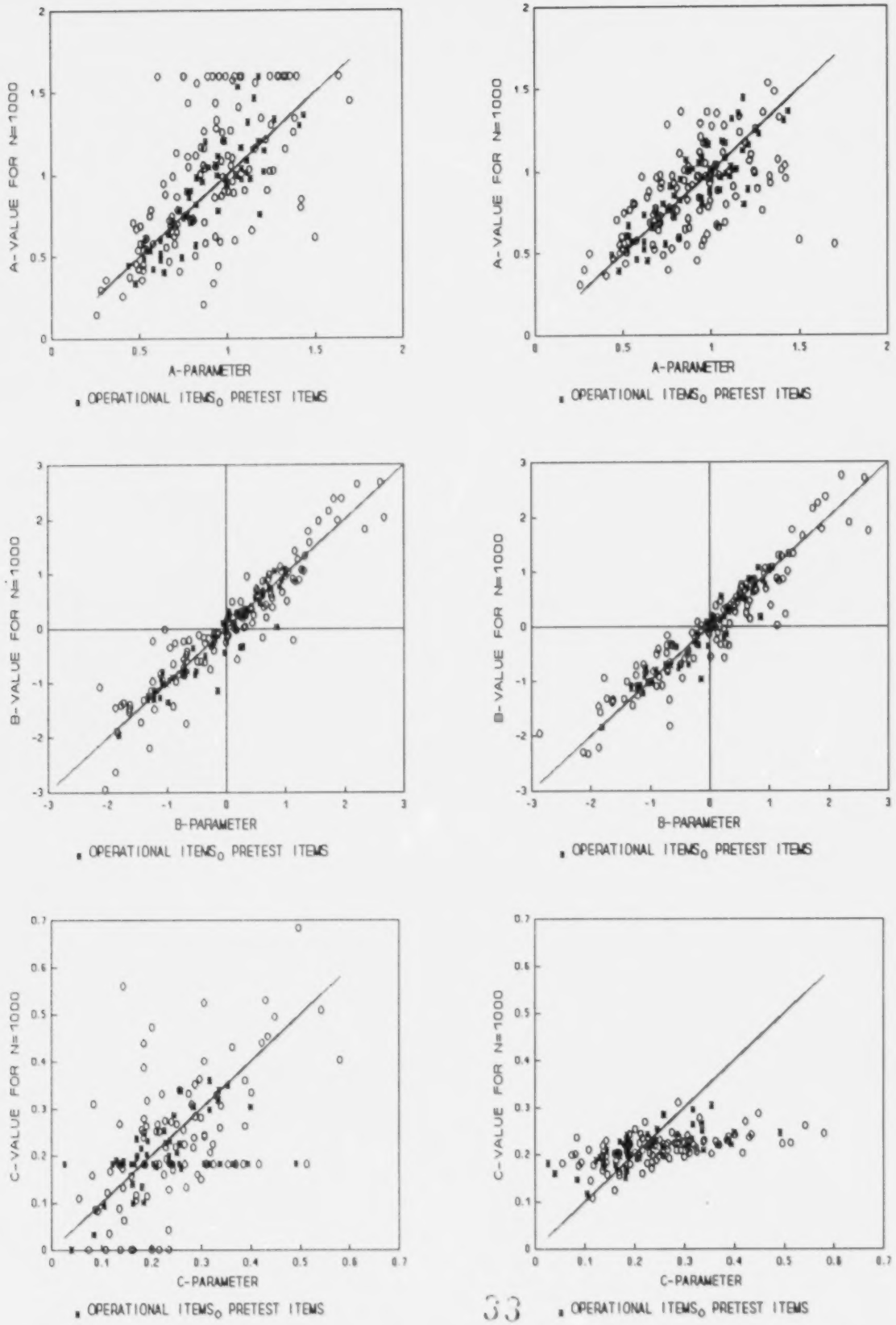


Figure 6: Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=1000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

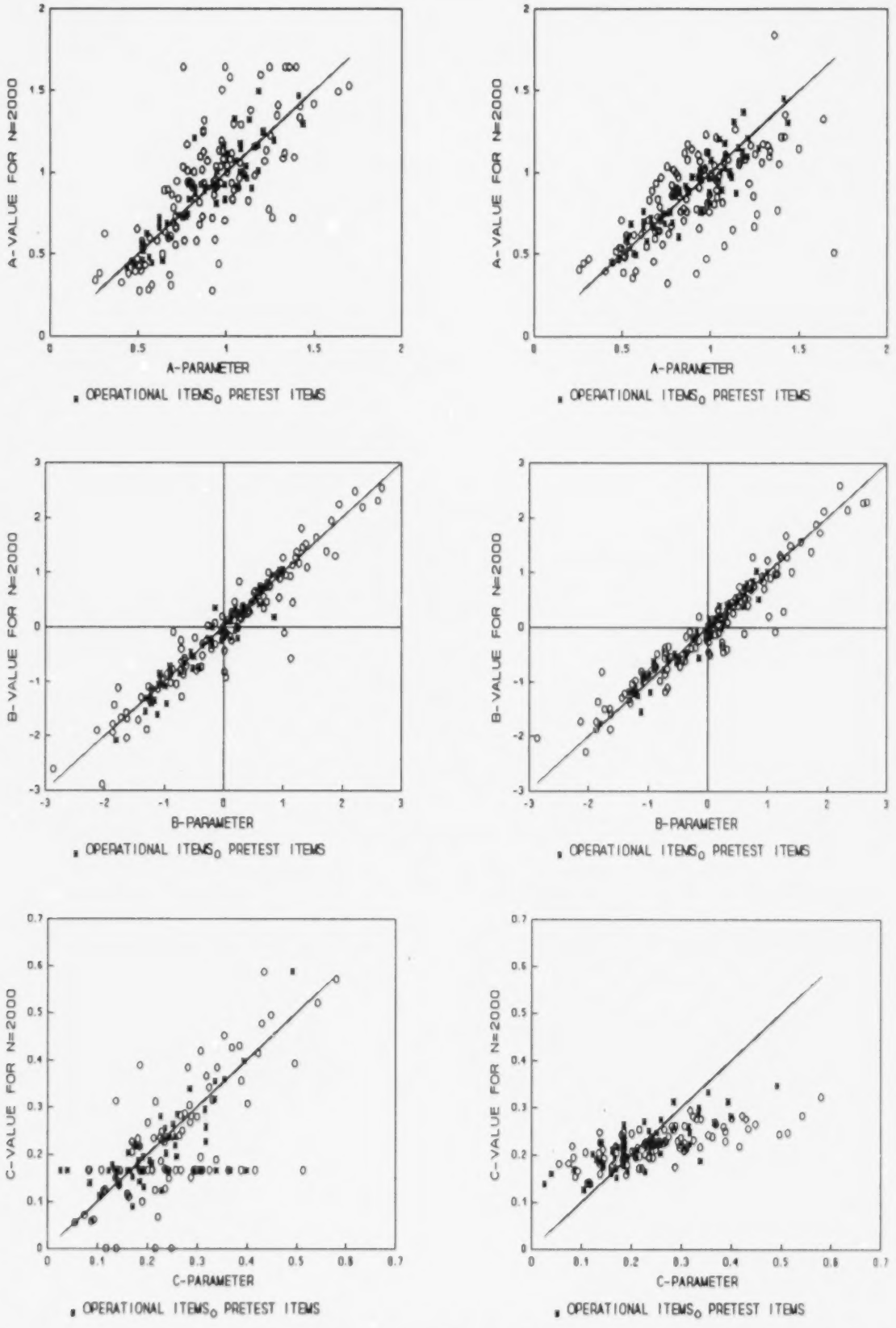


Figure 7: Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=2000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

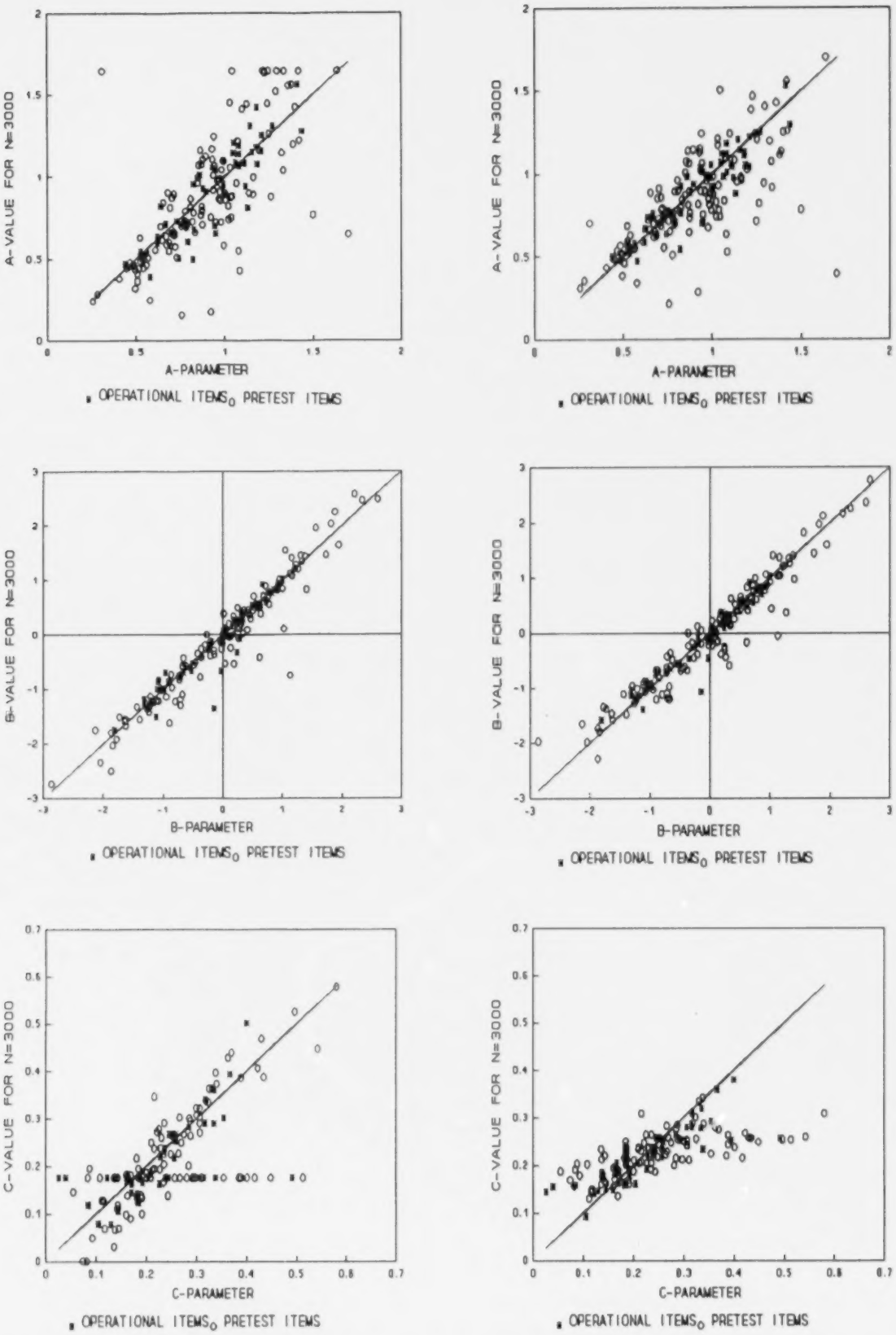


Figure 8: Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=3000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

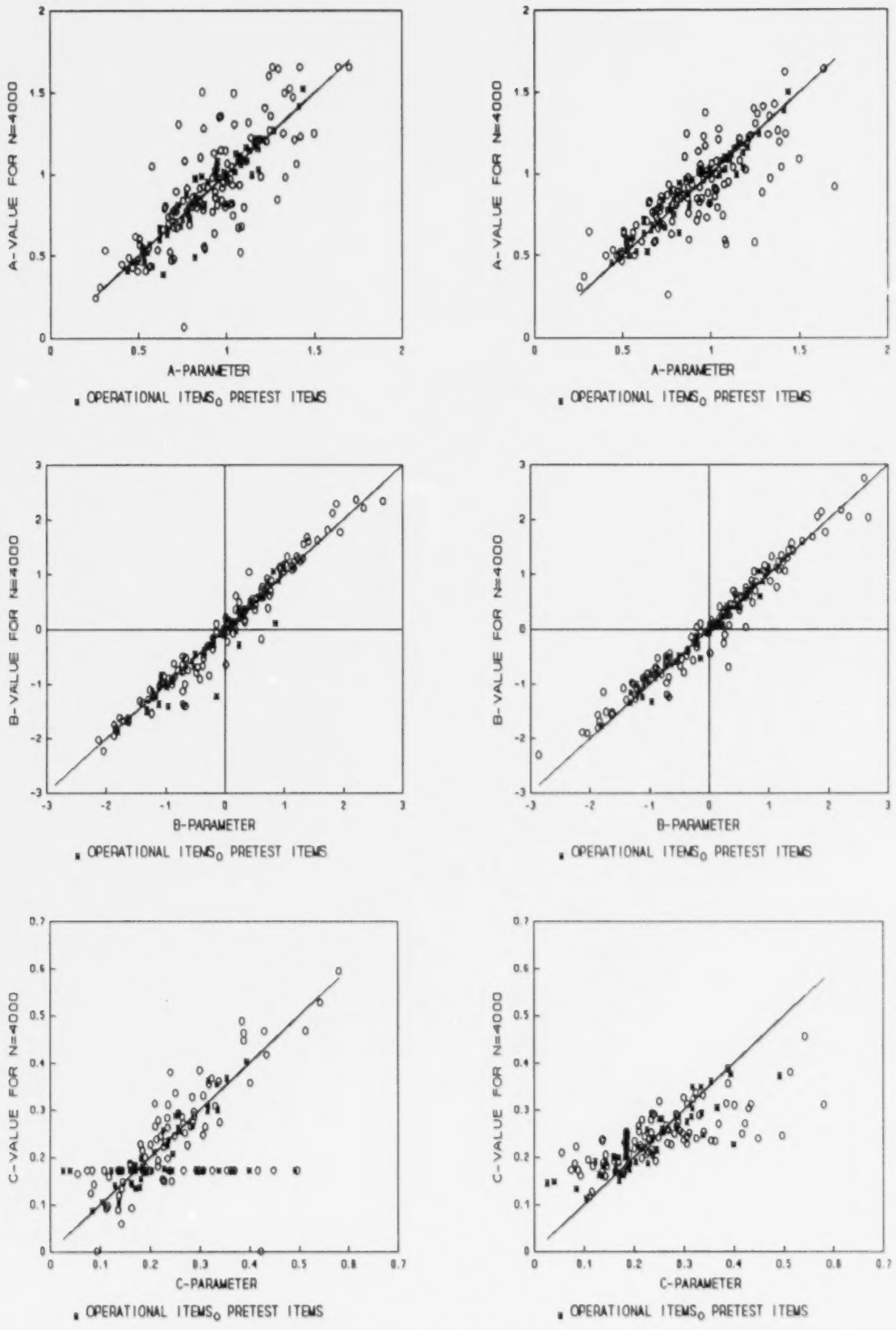


Figure 9: Plots of the Simulation I A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=4000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

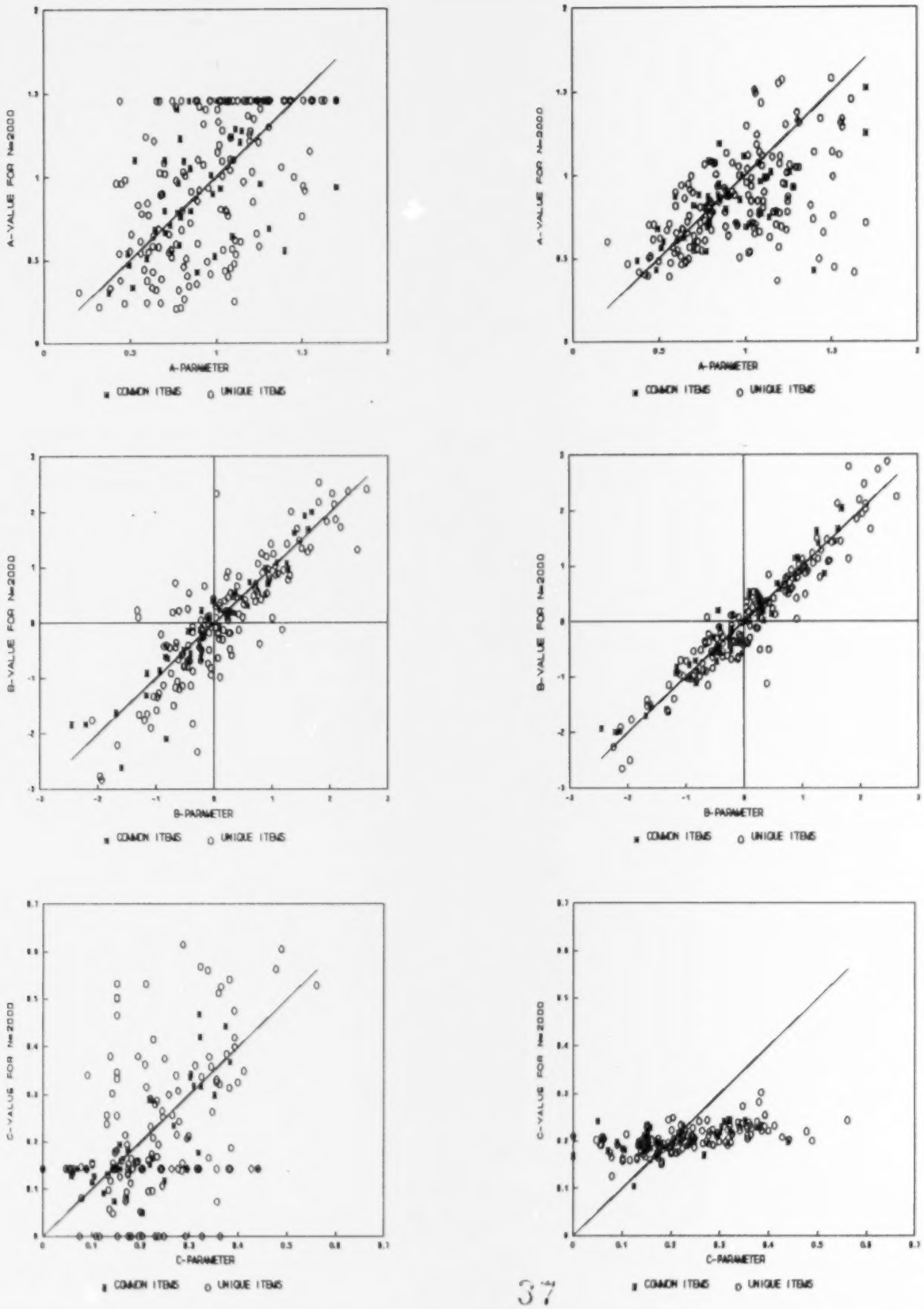


Figure 10: Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=2000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

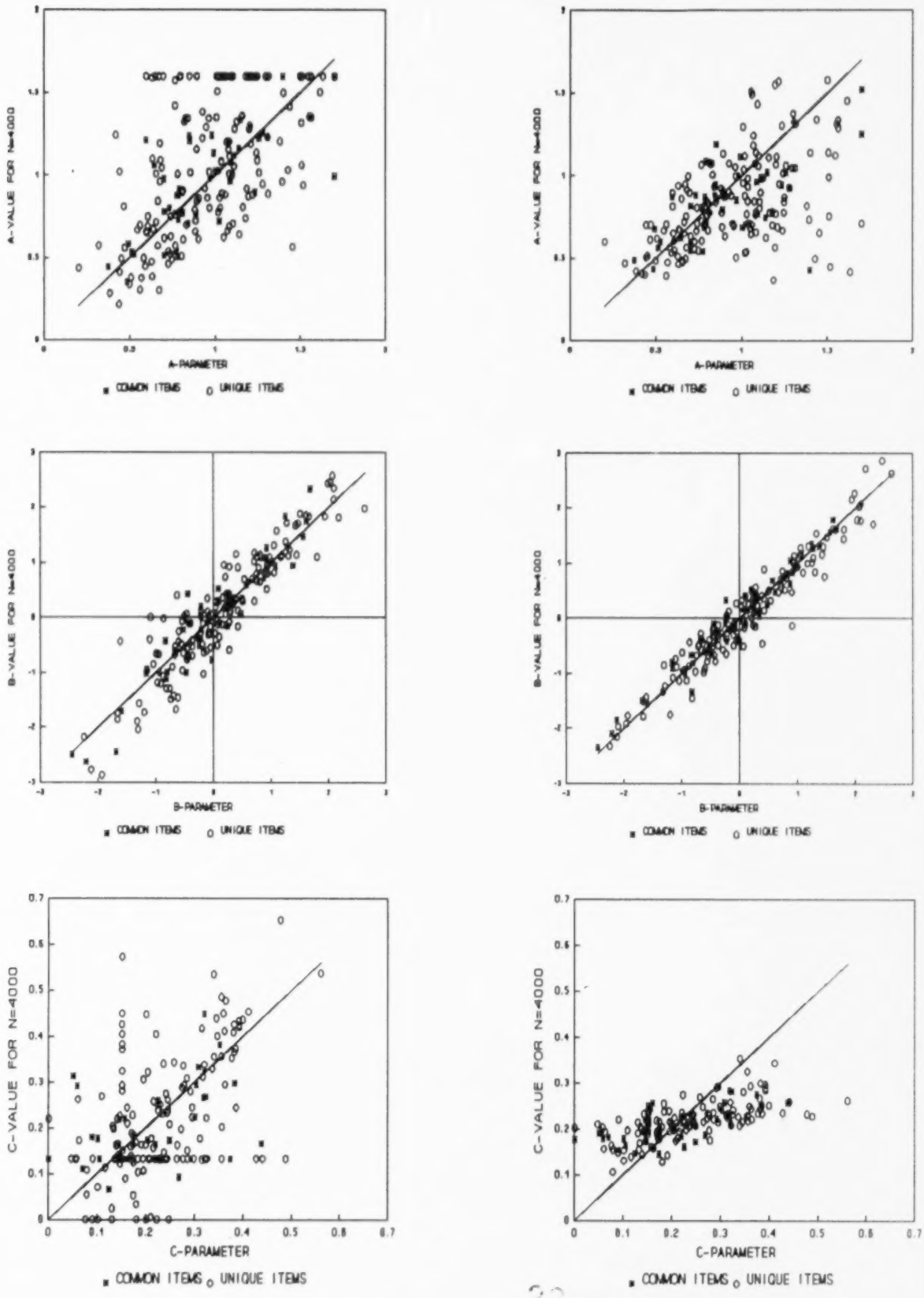


Figure 11: Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=4000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

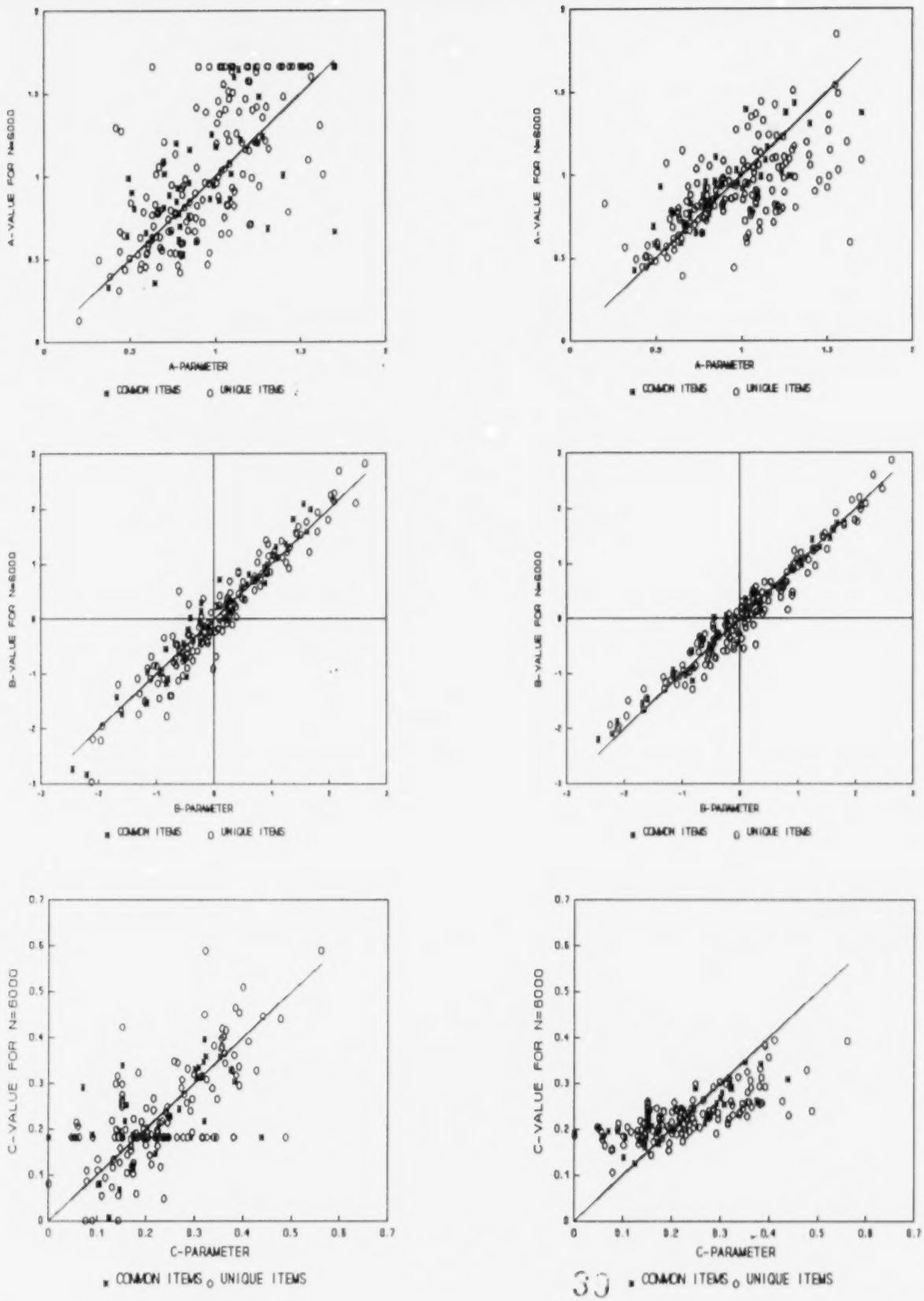


Figure 12: Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=6000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

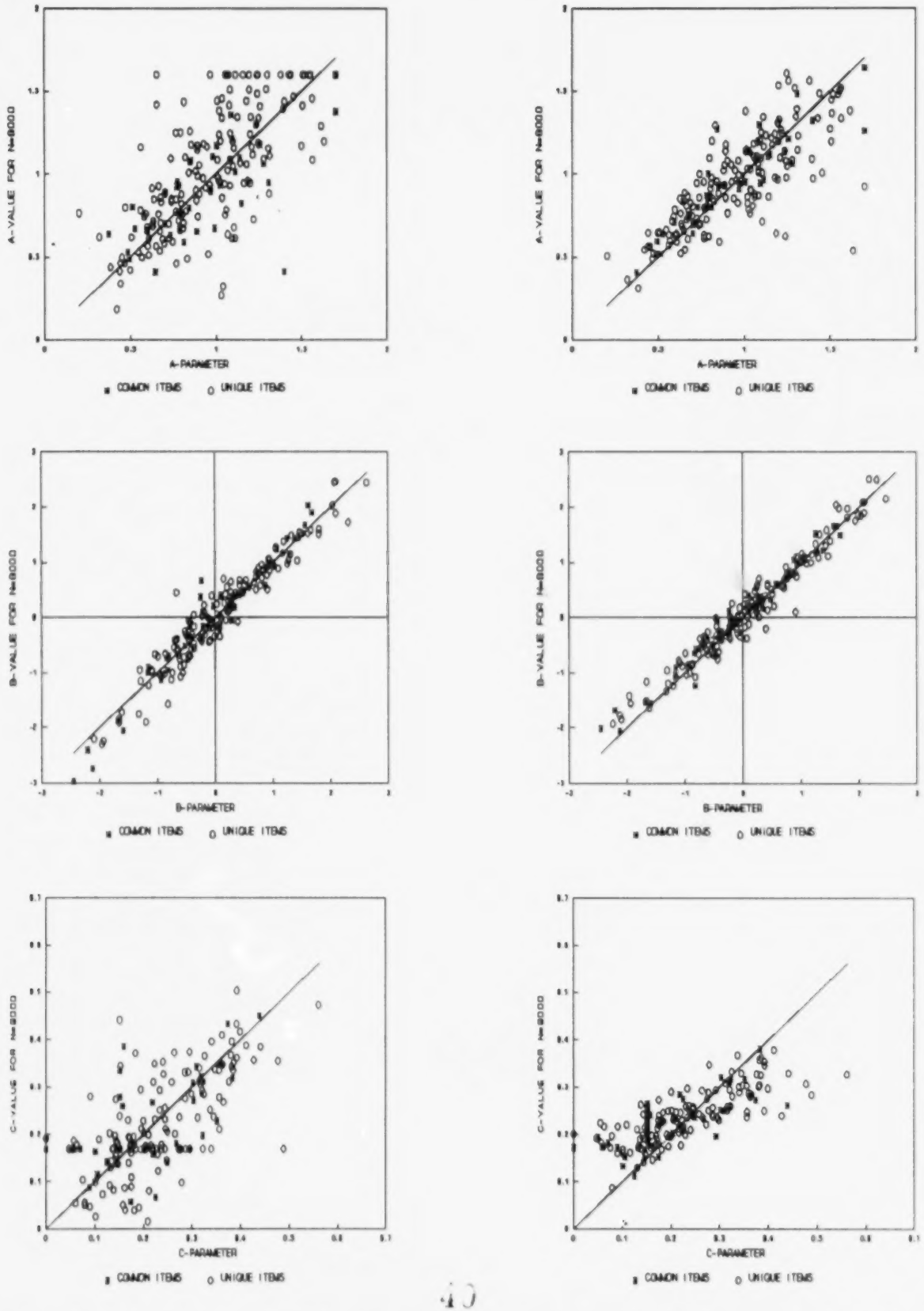


Figure 13: Plots of the Simulation II A-, B-, and C-Estimates for LOGIST (Left Hand Plots) and BILOG (Right Hand Plots) for the N=8000 Runs Vs. the Generating Parameters (All Estimates on the Scale of the Generating Parameters)

TOEFL Technical Report #7

RR-93-59

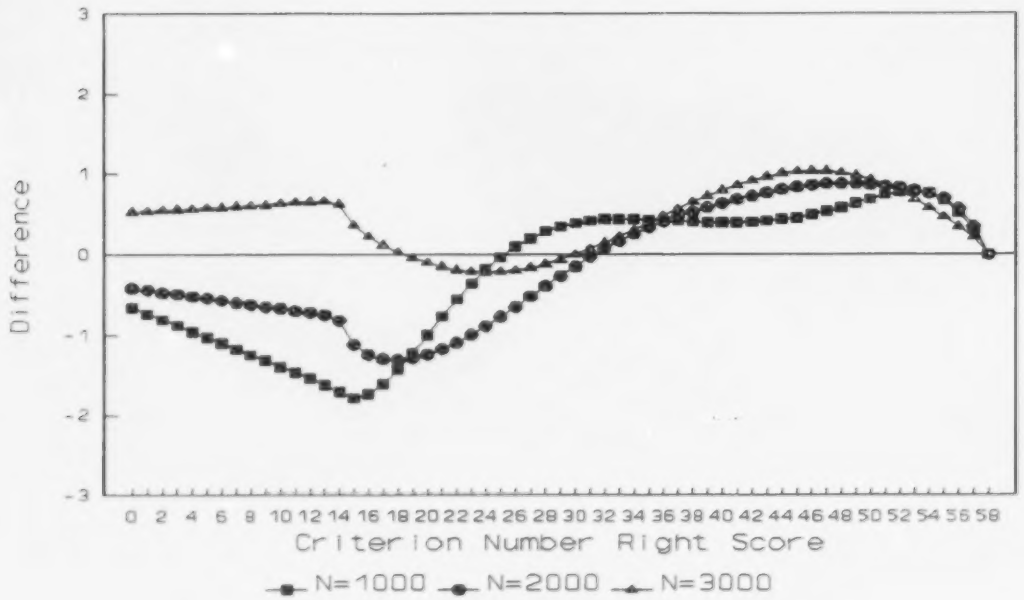
ERRATUM NOTICE

The attached three pages, Figures 14, 15, and 16,

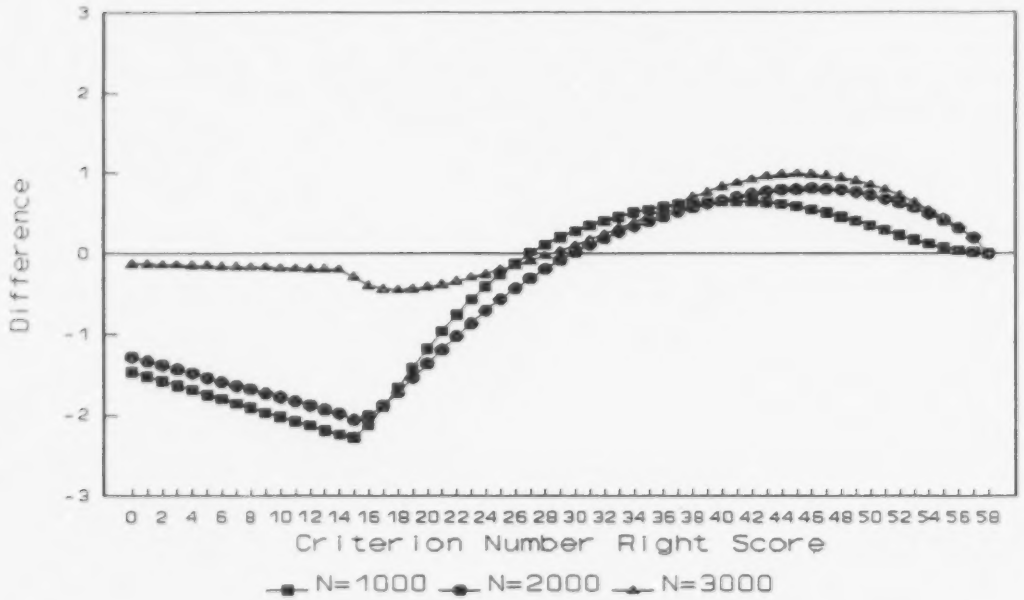
were not included in the original report.

TOEFL Program Office

January 1994

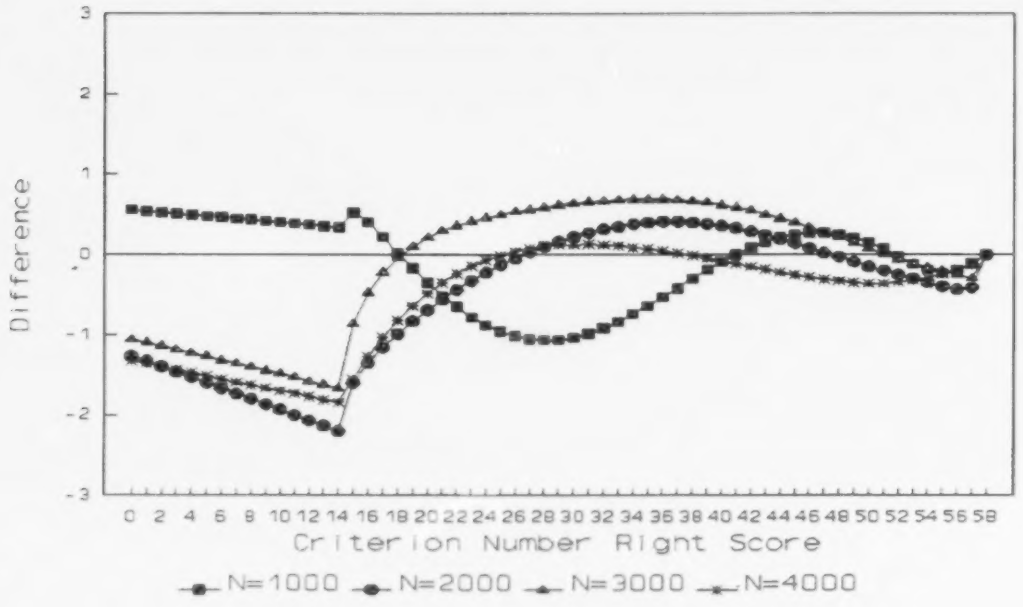


a) Equating Differences Based on LOGIST

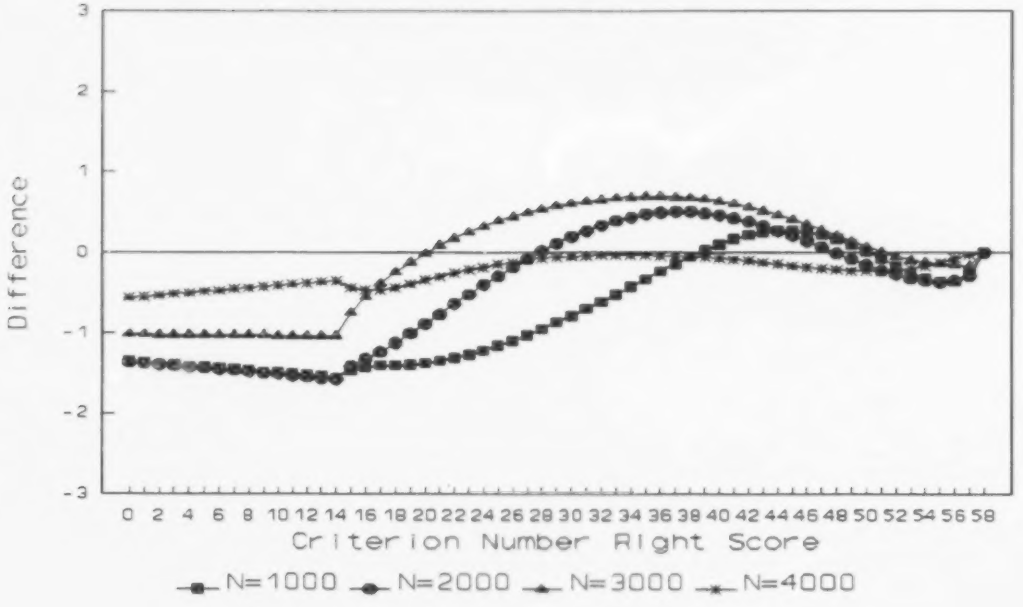


b) Equating Differences Based on BILOG

Figure 14: Equating Differences Between the BILOG and LOGIST Equated True Scores and the Criterion Scores based on the N = 4000 Samples - Real Data

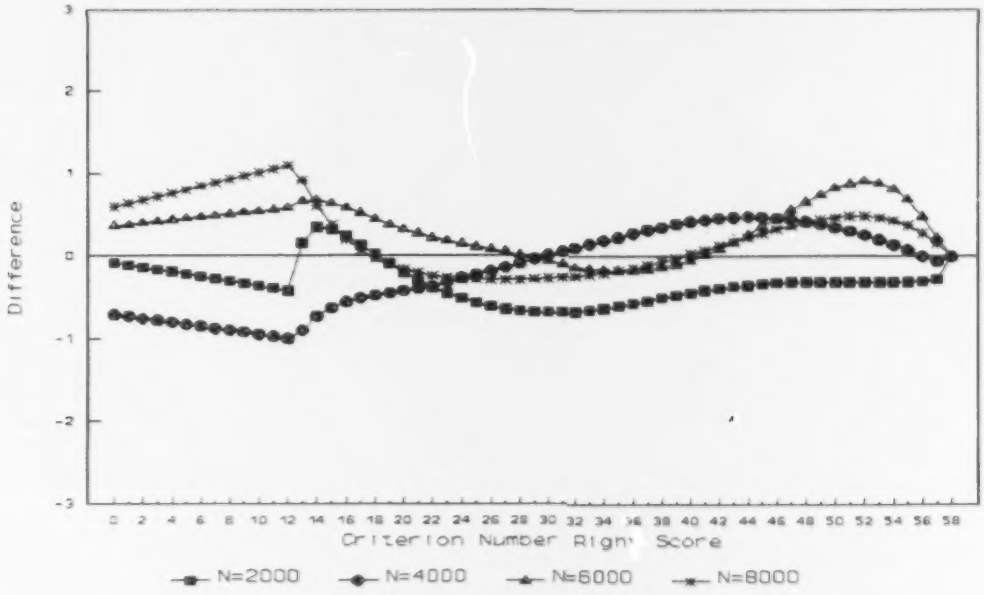


a) Equating Differences Based on LOGIST

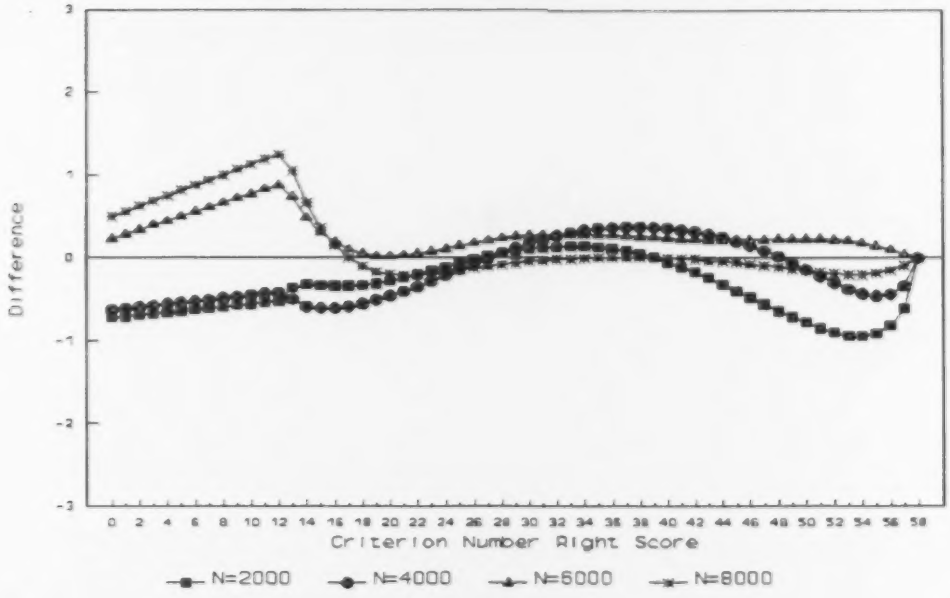


b) Equating Differences Based on BILOG

Figure 15: Equating Differences Between the BILOG and LOGIST Equated True Scores and the Simulating True Scores - Simulation I



a) Equating Differences Based on LOGIST



b) Equating Differences Based on BILOG

Figure 16: Equating Differences Between the BILOG and LOGIST Equated True Scores and the Simulating True Scores - Simulation II

Appendix A: LOGIST and BILOG Specifications

The following specifications for LOGIST and BILOG were used for the Real Data Case, Simulation I, and Simulation II calibrations.

LOGIST

1. The maximum value for the a-parameter was set to 1.7 and the initial value was set to 0.95.

2. The θs were restricted to the range of -7 to +7.

3. Individual c-parameters were estimated only for items with the value $b - 2/a > -2.5$.

4. Examinees with zero or perfect scores were removed from the estimation.

5. The maximum number of stages for one LOGIST run was set to 50 for the Real Data Case and Simulation I, and 70 or 90 for Simulation II.

6. The convergence for the criterion function was not checked for two stages.

7. The maximum number of CPU seconds for one LOGIST run was set to 9600 for the Real Data Case, 9900 for Simulation I, and 9999 for Simulation II.

8. The defaults were used for the rest of the LOGIST parameters.

BILOG

1. The θ distribution was held fixed as a standard normal distribution. The number of quadrature points of the θ distribution was 20, and the range of the quadrature points was between -4 and 4. The trim factor of the distribution was chosen to be 6.0.

2. In order to be consistent with LOGIST, the omits were treated as fractionally correct, which is equal to the inverse of the number of responses for each item.

3. The transformations of the item difficulties (P+) and biserial correlations were used as the initial estimates for the intercept parameters ($-a_i b_i$) and the slope parameters (a_i), respectively. The initial estimate of the pseudo-guessing parameter was chosen to be 0.2.

4. The number of E-M cycles was 30 for the Real Data Case and Simulation I, and 40 for Simulation II. The number of Newton-Raphson cycles was 3. The convergence criterion was 0.005.

5. The defaults were used for the rest of the BILOG parameters.

The choice of parameters for each program was made on the basis of some initial trial runs and the experiences of the authors in using the two programs. For LOGIST, the parameters were the same as those used in operational TOEFL calibrations with the exception of CRITFIXC, which was changed to -2.5 from -3.5 because of the small sample sizes used in the study. For BILOG, the weights of the priors were also chosen because of the small sample sizes used in the study. The choices affecting estimation were based in part on advice from ETS Research Scientists with operational experience using BILOG as part of the National Assessment of Educational Progress (NAEP).

Appendix B: Summary Statistics for the Data Sets Used in the Study

Table B.1
Score Means and Standard Deviations (SD) for the 58 Operational Items
In the Real Data Case and Simulation I

	Real Data		Simulation I	
	Mean	SD	Mean	SD
Sample Size = 1000				
Group 1 (N=250)	36.76	11.21	37.46	10.93
Group 2 (N=250)	37.22	9.85	37.98	11.89
Group 3 (N=250)	36.15	11.80	36.99	11.02
Group 4 (N=250)	36.24	11.18	37.10	11.89
Sample Size = 2000				
Group 1 (N=500)	37.16	11.42	38.10	11.06
Group 2 (N=500)	37.90	10.57	37.18	11.50
Group 3 (N=500)	36.23	11.25	37.54	10.85
Group 4 (N=500)	36.19	11.81	36.94	11.23
Sample Size = 3000				
Group 1 (N=750)	36.54	11.39	36.99	11.18
Group 2 (N=750)	36.74	10.60	37.09	11.14
Group 3 (N=750)	36.86	11.26	36.72	11.15
Group 4 (N=750)	36.26	11.24	37.28	11.43
Sample Size = 4000				
Group 1 (N=1000)	37.72	10.81	37.56	11.35
Group 2 (N=1000)	37.26	10.53	37.41	10.85
Group 3 (N=1000)	37.28	11.29	36.58	11.55
Group 4 (N=1000)	36.41	11.62	37.27	11.32

Table B.2
Score Means and Standard Deviations (SD) for the 30 Pretest Items
In the Real Data Case and Simulation I

	Real Data		Simulation I	
	Mean	SD	Mean	SD
Sample Size = 1000				
Group 1 (N=250)	18.72	5.17	18.88	5.36
Group 2 (N=250)	16.09	5.35	16.60	6.10
Group 3 (N=250)	18.27	5.92	18.34	6.07
Group 4 (N=250)	18.62	5.00	18.86	5.59
Sample Size = 2000				
Group 1 (N=500)	18.89	5.63	19.10	5.44
Group 2 (N=500)	16.49	5.41	16.42	5.81
Group 3 (N=500)	18.15	5.66	18.87	5.62
Group 4 (N=500)	18.72	5.40	18.98	5.10
Sample Size = 3000				
Group 1 (N=750)	18.62	5.36	18.91	5.29
Group 2 (N=750)	16.27	5.33	16.23	5.64
Group 3 (N=750)	18.58	5.58	18.67	5.57
Group 4 (N=750)	18.65	5.17	19.01	5.25
Sample Size = 4000				
Group 1 (N=1000)	18.77	5.15	18.86	5.41
Group 2 (N=1000)	16.23	5.22	16.20	5.51
Group 3 (N=1000)	18.86	5.61	18.53	5.68
Group 4 (N=1000)	18.55	5.45	19.06	5.33

Note. Group 2 means are consistently lower across sample size because this pretest set was more difficult than the others.

Table B.3
Score Means and Standard Deviations (SD) for the 30 Items
In Simulation II

	Mean	SD	Mean	SD
	Sample Size = 2000 (N = 250)		Sample Size = 4000 (N = 500)	
Group 1	18.70	5.27	18.10	5.73
Group 2	17.46	5.68	17.11	5.45
Group 3	20.20	6.10	19.58	5.68
Group 4	18.00	5.81	18.70	5.63
Group 5	17.80	5.48	17.16	5.63
Group 6	19.02	5.35	18.50	5.86
Group 7	17.18	5.63	17.03	5.71
Group 8	17.50	5.66	17.84	5.91
	Sample Size = 6000 (N = 750)		Sample Size = 8000 (N = 1000)	
Group 1	18.22	5.58	18.20	5.73
Group 2	16.88	5.37	16.88	5.36
Group 3	20.45	5.54	19.60	5.84
Group 4	18.03	6.06	18.39	5.93
Group 5	17.37	5.70	17.74	5.71
Group 6	19.33	5.57	18.78	5.41
Group 7	17.47	5.90	17.17	5.86
Group 8	17.54	5.51	17.52	5.66

Note. Group 3 means are consistently higher across sample size because this item set was easier than the others.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Broch, E., & McKinley, R. L. (April 1991). An evaluation of the item pool calibration sample size requirements for computerized adaptive testing for the NTE successor stage I test. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Dygert, E. H. (1989). PC-LOGIST, LOGIST Version 6.0. Menu System User's Guide. Princeton, NJ: Educational Testing Service.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small data sets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement, 15, 279-291.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1986). Bayes modal estimation in item response model. Psychometrika, 51, 177-195.

Mislevy, R. J., & Bock, R. D. (1984). BILOG I: Maximum likelihood item analysis and test scoring with binary logistic models. Mooresville, Indiana: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1989). PC-BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, Indiana: Scientific Software, Inc.

Mislevy, R. J., & Stocking M. L. (1989). A consumers guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

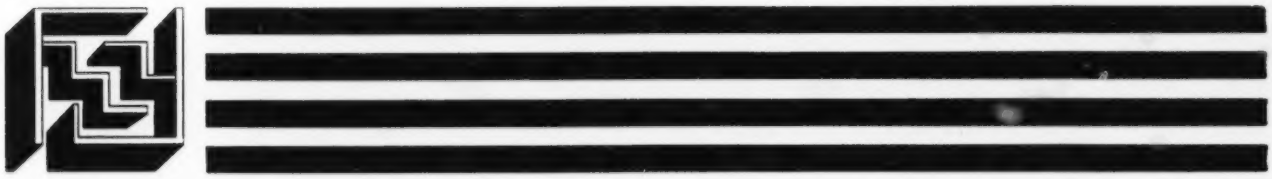
Way, W. D., & Reese, C. M. (1991). An investigation of the use of simplified IRT models for scaling and equating the TOEFL test (TOEFL Technical Report). Princeton, NJ: Educational Testing Service.

Way, W. D., Twing, J. S., & Ansley, T. A. (April 1988). A comparison of vertical scalings with the three-parameter model using LOGIST and BILOG and two different calibration procedures. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). LOGIST user's guide. Version 6.0. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

TOEFL is a program of
Educational Testing Service
Princeton, New Jersey, USA



57906 • Y123M.5 • 275599 • Printed in U.S.A.

BEST COPY AVAILABLE